Enabling data download using gsiftp
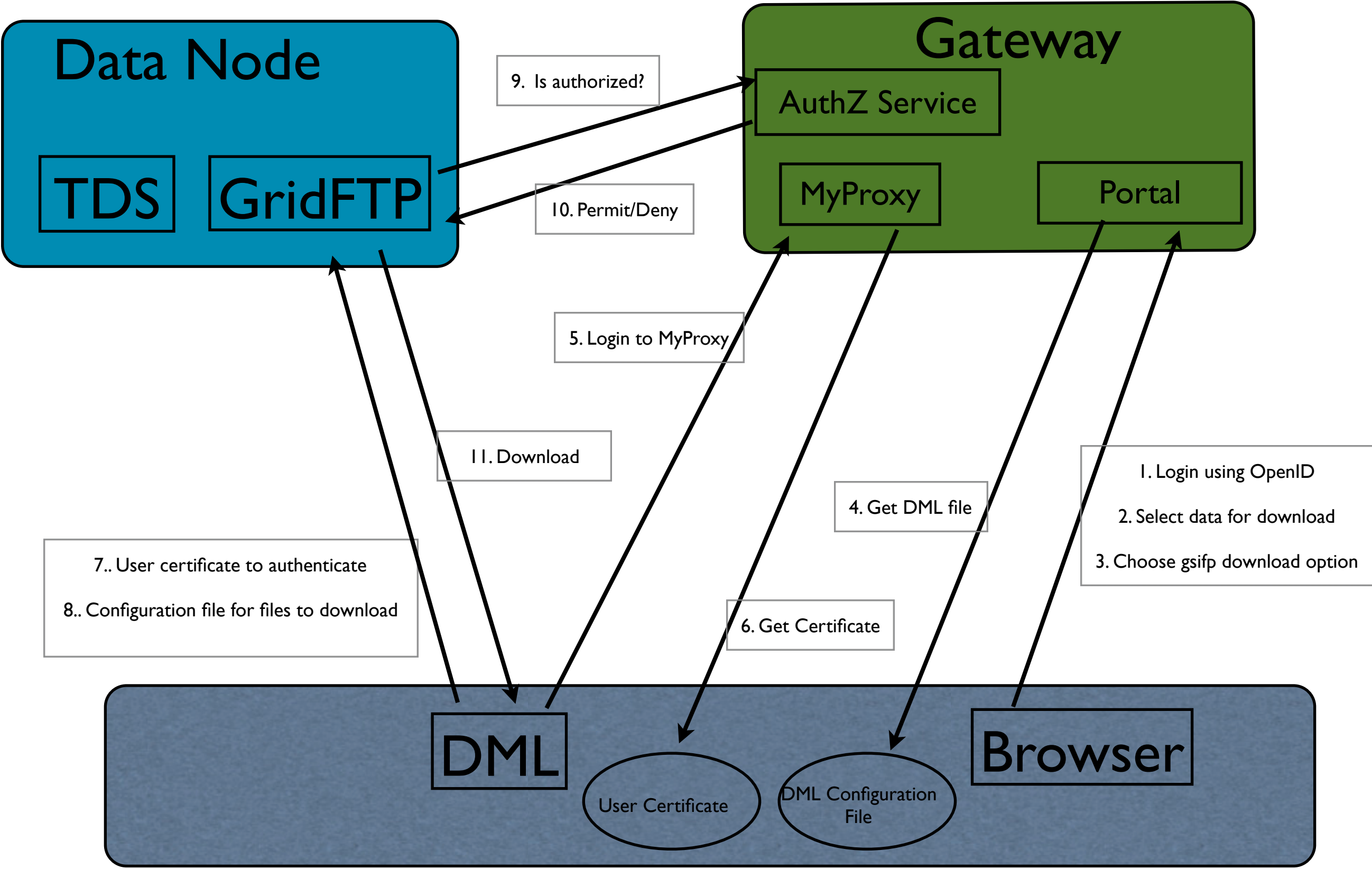
* Working to enable use of GridFTP to move data to the end user.
* Be able to leverage performance, retries, checksum and other interesting features for data movement to the user.
* Context of this discussion is only user download of data, nothing to do with Bulk Data Movement

Data download scenario

Data Node

TDS | GridFTP

Gateway

AuthZ Service

MyProxy | Portal

9. Is authorized?

10. Permit/Deny

5. Login to MyProxy

11. Download

7.. User certificate to authenticate

8.. Configuration file for files to download

4. Get DML file

1. Login using OpenID

2. Select data for download

3. Choose gsifp download option

6. Get Certificate

DML | Browser

User Certificate | DML Configuration File

GridFTP Server

- Authorization

  - Parse token from URL

  - Present to authorization service (already exists) to authorize request

  - Permit/deny enforcement

  - Callout a C Hessian Client to talk to authorization service

    Status from Neill

Additions required on GridFTP server:
* authorization of the requests that come in
* token embedded in the URL will be pulled out and sent to a configured authorization service
* if the service returns permit, data is transfered. if deny is returned, an error is thrown to the user
* Note that in this mechanism, given a gsiftp URL with token in a said time frame, any one can download the file.
That is if Neill shares his URL with Rachana, she can download the file. It is no worse off from how it is done in
TDS today, but something to think about in future.

Gateway & Publishing

- THREDDs catalog needs to be populated with gsiftp URLs

    - First phase, assuming out of band knowledge about presence of GridFTP server and endpoint

- Downloadable DML configuration file from Gateway

Status from Bob & Luca

Publishing:
* Currently publisher populates with http download URLs
* gsiftp requires that the URLs be absolute path and not all data nodes will have GridFTP
* gsiftp URLs will need to be added to catalog for Gateway to use.
Gateway:
* will allow user to choose http or gsiftp download, if latter is available
* in gsiftp case at-least, the download will be configuration file consumed by DML
* No changes required on browsing and other gateway functions.

Data Mover Lite

- Integrate MyProxy client to be able to log on to MyProxy

- Integrate MyProxy bootstrap client functionality

- Use downloaded user credential to start download

- Integrate automatic download (Java Web Start) for ease of deployment

Status from Alex

* Used as primary client for gsiftp download
* Should download MyProxy trust certificates in anonymous mode to bootstrap client
* Should allow user to enter username/password and choose MyProxy sever (or intelligent guess) and get user's credential
* Use the credential to start transfer based on configuration file
* Allow use of existing credential as an option
* Integrate with Java Web Start, will allow clients to click on a link in browser and application will be self contained to install everything

Next Steps

- Explore and enable direct data download to analysis tools use case

- Smart download - gateway cache or direct data download

- Discovery service for ESG federation

- GridFTP Performance Tuning

- Investigate data download policy

- Authorization improvements

    - prevent sharing of download permissions

    - alternative to callouts?

These are just ideas at this point, need to collect use cases and discuss.
* Client has analysis tool on his machine and out of band knowledge of data to download. Gateway is not used and data is pulled by analysis tool
* Smart download (Arie suggested): way to determine if it makes more sense to allow user to just download form gateway cache rather than direct data node download
* Discovery Service: lots of endpoints and common services. Useful to define a discovery service that has information about GridFTP servers, authorization services and other useful endpoints. Useful in production systems (e.g TG, BIRN)
* GridFTP performance tuning: gather usage statistics to tune the server and client. For example, leverage performance tuning for lots of small files using pipelining.
* Ascertain data download policy, limits on size of download
* Authorization today allows URLs to be shared, use cases for this and options. Alternative to callout, pushing assertions to the server that inform the server of the rights rather than server going to a authorization service. Maybe both options – need to gather use cases for this.